不确定性数据中基于 GSO 优化 MF 的模糊关联规则挖掘方法 *

章武媚 [1,2]，董 琼 [1]

(1. 美国纽约州立大学，美国 纽约州 奥斯威戈 13126; 2. 浙江同济科技职业学院，杭州 310023)

摘 要：针对不确定性数据中模糊关联规则的挖掘问题，提出一种基于群搜索优化(GSO)算法优化隶属度函数(MF)的模糊关联规则挖掘方法。首先，将不确定性数据通过三元语言表示模型进行表示；然后，给定一个初始 MF，并以最大化模糊项集支持度和语义可解释性作为适应度函数，通过 GSO 算法的优化学习获得最佳 MF；最后，根据获得的最佳 MF，利用改进型的 FFP-growth 算法来从不确定数据中挖掘模糊关联规则。实验结果表明，该方法能够根据数据集自适应优化 MF，以此实现从不确定数据中有效地挖掘关联规则。

关键词：模糊关联规则挖掘；不确定数据；隶属度函数；群搜索优化算法；FFP-growth 算法

中图分类号：TP311 doi: 10.3969/j.issn.1001-3695.2018.01.0076

Fuzzy association rules mining method based on GSO optimization MF in uncertainty data

Zhang Wumei[1, 2], Dong Qiong[1]

(1. *State University of New York*, *New York*, *Oswego*, *America* 13126, *USA*; 2. *Zhejiang Tongji Vocational College of Science & Technology*, *Hangzhou* 310023, *China*)

Abstract: In order to solve the problem of mining fuzzy association rules in uncertainty data, this paper proposed a new method of mining fuzzy association rules based on optimization of membership function (MF) by group search optimization (GSO) algorithm. Firstly, it represented the uncertainty data by the 3-tuples linguistic representation model. Then, given an initial MF, it obtained the best MF by optimizing learning of GSO algorithm with maximum support of fuzzy itemsets and semantic interpretability as a fitness function. Finally, it used the improved FFP-growth algorithm to mine the fuzzy association rules from the uncertain data according to the best MF obtained. Experimental results show that this method can adaptively optimize MF based on data set, so as to effectively mine association rules from uncertain data.

Key words: fuzzy association rule mining; uncertainty data; membership function; group search optimization algorithm; FFP-growth algorithm

0 引言

关联规则表示数据库项目之间的依赖关系，已广泛应用于众多领域中，如市场分析、入侵检测、诊断决策以及电信领域[1]。然而真实的数据库通常具有不确定性。不确定性包含概率性和不完整性数据[2]。如何有效地从不确定性数据库中挖掘事物之间的关系已成为一个主要研究方向。

近年来，许多研究人员提出了挖掘模糊关联规则(fuzzy association rule，FAR)的方法[3,4]，用来扩展可能的关系类型，以便于在语言学方面对规则进行解释[5]。从不确定性数据库中挖掘 FAR 需要设计能够处理不精确数据的算法。例如，文献[6]使用不确定性的可能表示法从不确定数据中挖掘 FAR，其可以应对数据以区间和模糊值作为输入。但是该方法假定隶属度函数(membership function，MF)是事先已知的，但模糊关联规则挖掘的性能与 MF 的位置密切相关，给定的 MF 不能适应环境的变化，且事先决定最合适的 MF 很难做到。

为此，研究一种能够从不确定数据库中进行自适应学习的方法，获得一组合适的 MF 来挖掘 FAR，将会具有重要的意义。目前，自动优化 MF 的方法主要分为两种，一种是基于神经网络学习的优化方法，如常用的 RBF 神经网络[7]。然而，基于神经网络的方法需要大量的可靠训练样本，在实际应用中，获得不同 MF 下的规则挖掘样本比较困难。第二种是利用一些智能搜索算法来调整模糊系统中的 MF[8,9]，如，蚁群算法、遗传算法、粒子群算法等。在多种智能优化算法中，群搜索优化(group search optimize，GSO)算法具有很强的全局搜索能力，对于函数优化问题有明显优势。但是，若直接采用这些算法从搜索空间中寻找最优 MF，优化时间较长。对此，可通过一些新的语言规则表示模型来降低优化时间。例如，文献[10]提出了一种新的语

收稿日期：2018-01-31；**修回日期：**2018-03-12 **基金项目：**国家自然科学基金资助项目（166223123）；浙江省教改项目（jg20160405）

作者简介：章武媚（1971-），女，浙江永康人，教授，硕士，主要研究方向为计算机应用技术研究（zhangwumei71@126.com）；董琼，女，教授，博士，主要研究方向为超网络理论及应用.

言规则表示模型来进行 MF 的调整。这个新的模型是基于三元语言表示的，它通过仅考虑两个参数来支持语言学术语的横向位移和支持度变化。该方法能够通过调整 MF 来获得高度的数据覆盖，并且减少经典挖掘方法的搜索空间。

鉴于上述分析，本文将三元语言表示法与 GSO 算法相结合，利用三元语言表示法来降低算法搜索空间，使 GSO 算法能够快速寻优 MF。为此提出了一种新的不确定数据模糊关联规则挖掘算法，从不确定数据库中学习合适的 MF，并挖掘高效的 FAR。本文方法主要的创新点如下：

a) 为了能够获得最合适的 MF，本文基于三元语言表示模型和 GSO 算法进行 MF 的学习优化，最大化模糊支持度和可解释性度量，以此用来减少搜索空间并保持 MF 的语义解释能力。

b) 为了减少挖掘时间，本文还提出了一种新的数据挖掘(DM)算法：不确定模糊频繁模式增长算法(uncertair- FFP-growth，*U-FFP-growth*)。即利用学习到的 MF，从不确定数据库中有效地挖掘 FAR。*U-FFP-growth* 算法是模糊频繁模式增长算法(FFP-growth)的扩展，用来从不确定数据中挖掘模糊关联规则且不需要生成候选项集。

1　提出的模糊规则挖掘方法框架

本文提出了一种从不确定数据中基于隶属函数(MF)学习的模糊关联规则(FAR)挖掘算法，称为 *U-MFL-FAR*。

提出的方案包括两个阶段：首先，基于三元语义表示模型和 GSO 算法进行优化学习来获得 MF，最大化模糊项集的模糊支持度和 MF 的解释性；然后，对 FFP-growth 算法进行扩展，使其能够基于优化后的 MF，从不确定数据中挖掘有用的 FAR。提出方法的框架如图 1 所示。

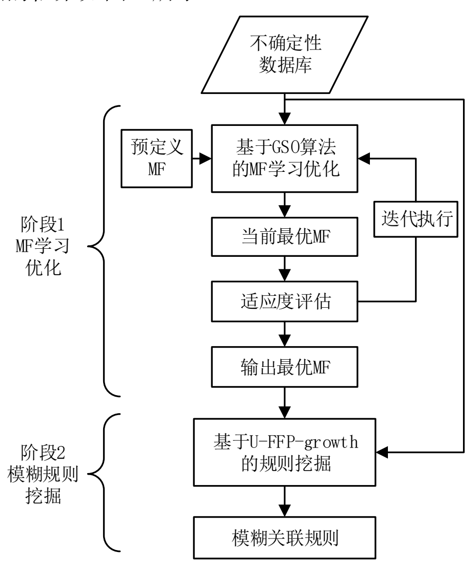


图 1　提出的不确定性数据模糊关联规则挖掘方法框架

2　三元组语言表示模型

在模糊系统的框架中，主要采用的是三角形 MF。可通过优化方法对其中三个参数进行细化，这三个参数确定了与数据标签相关的 MF[11]。然而在存在多变量问题的情况下，MF 之间的依赖关系和三个参数之间的依赖关系会导致优化模型需要处理非常复杂的搜索空间，影响了优化性能[12]。

在文献[10]中提出了一种基于三元组语言表示方法的新的规则表示模型。该方案考虑两个参数 α 和 β，分别表示标签的横向位移和支持度变化。这样，每个语言术语可以由一个三元组 (s,α,β) 表示。其中 α 是区间[-0.5,0.5)内的一个数，这使得可以对标签进行横向位移，直到达到两个相邻标签距离的 50%；β 也是一个在[-0.5,0.5)范围内的数字，这使得可以增加或减少标签的支持度，直到其为原始大小的 50%。例如，图 2 显示了一个由三元组表示的标签 $s_2'=(s_2,-0.3,-0.25)$ 以及相应 MF 的横向位移和支持度变化。

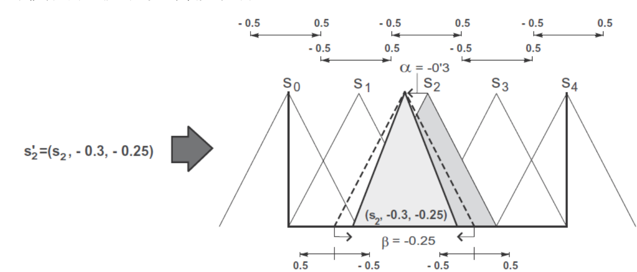


图 2　标签 s_2 的横向位移和支持度变化

这种新的规则表示模型允许通过学习各自的横向位移和支持度变化来调整 MF，从而可以有效减小搜索空间。

考虑一个简单的挖掘问题，其中包含与语言术语有关的两个变量(*Age* 和 *Height*)。基于这个定义，经典的 FAR 和三元组模糊语言表示的 FAR 分别为：

经典 *FAR*：if *Age* 是小的，then *Height* 是小的。

三元组模糊语言表示的 FAR：if *Age* 是 $(Low,0.1,0.1)$，then *Height* 是 $(Low,0.1,-0.1)$。

3　基于 GSO 的隶属函数优化

3.1　GSO 算法

对于求解优化问题，启发式智能算法最为有效。其中，GSO 算法包含三个操作，即发现者操作、搜索者操作和游荡者操作。在迭代过程中，将具有最佳适应度值的成员选为发现者。将适应度值高于阈值的多个成员选为搜索者，将适应度值低于阈值的多个成员选为游荡者。

a)发现者操作：

发现者操作过程中，动物旋转感官受体从环境中捕获信息。在 s 维搜索空间中，第 z 个搜索回合(迭代)的第 i 个成员的位置表示为 $y_i^z\in R^s$，搜索角度表示为 $\lambda_i^z=(\lambda_{i1}^z,\ldots,\lambda_{i(s-1)}^z)\in R^{s-1}$，对应的搜索方向表示为 $F_i^z(\lambda_i^z)=(f_{i1}^z,\ldots,f_{is}^z)\in R^{s-1}$，其可以通过极坐标变换根据 λ_i^z 计算得到，表达式如下：

$$
\begin{aligned}
f_{i1}^z &= \prod_{p=1}^{s-1}\cos(\lambda_{ip}^z);\\
f_{ij}^z &= \sin(\lambda_{i(j-1)}^z)*f_{i1}^z;\\
f_{is}^z &= \sin(\lambda_{i(j-1)}^z);
\end{aligned}
\tag{1}
$$

假设在第 z 次迭代处的发现者位置为 y_p，那么发现者将会在当前位置选择三个不同的角度进行视觉扫描，即首先以零度扫描，然后向右边扫描，再向左边扫描。设定视觉的最大搜

索角度为 $\omega_{\max}$，视觉扫描的最大距离为 $d_{\max}$，表达式如下：

$$d_{\max}=\|U_i-L_i\|=\sqrt{\sum_{i=1}^{s}(U_i-L_i)^2} \tag{2}$$

其中：U_i 和 L_i 分别为设计变量取值范围的上界和下界。

那么，发现者通过扫描所发现的三个不同位置表示如下：

$$\begin{cases} y_{zero}=y_p^z+r_1d_{\max}F_p^z(\lambda^z) \\ y_{right}=y_p^z+r_1d_{\max}F_p^z(\lambda^z+r_2\dfrac{\omega_{\max}}{2}) \\ y_{left}=y_p^z+r_1d_{\max}F_p^z(\lambda^z-r_2\dfrac{\omega_{\max}}{2}) \end{cases} \tag{3}$$

其中：y_{zero} 表示零度扫描；y_{right} 表示右边扫描；y_{left} 表示左边扫描；$r_1\in R^1$ 为均值为 0、方差为 1 的正态分布随机数，$r_2\in R^{s-1}$ 为 $(0,1)$ 内的一个随机序列。

然后，计算发现者搜索到的三个新位置的适应度，并移动到具有最优适应度的位置。如果新位置都不如当前位置，则将其头转向一个新角度，表示如下：

$$\lambda^{z+1}=\lambda^z+r_2\gamma_{\max} \tag{4}$$

其中：$\gamma_{\max}$ 表示最大转向角。如果在 a 次迭代结束后，发现者没有找到一个更好位置，则停止搜索过程且保持不动，即

$$\lambda^{z+a}=\lambda^z \tag{5}$$

b)搜索者操作：

搜索者操作为跟随发现者，并在其周围附近区域进行搜索。在第 z 次迭代处，第 i 个搜索者根据发现者共享的位置信息执行区域搜索[13]，其位置更新如式（6）所示。

$$y_i^{z+1}=y_i^z+r_3(y_p^z-y_i^z) \tag{6}$$

其中：$r_3\in R_s$ 表示 $(0,1)$ 区间内的随机数。

c)游荡者操作：

游荡者操作仅为随机游走，并以此来探索新位置。如果将群中第 i 个成员选为第 z 次迭代的游荡者，则它将生成一个随机角度 λ_i，其表示如下：

$$\lambda_i^{z+1}=\lambda_i^z+r_2\gamma_{\max} \tag{7}$$

同样，也会选择一个随机距离，表示如下：

$$d_i=a\cdot r_1d_{\max} \tag{8}$$

然后，根据式（9）移向一个新位置：

$$y_i^{z+1}=y_i^z+d_iF_i^z(\lambda^{z+1}) \tag{9}$$

3.2 个体适应度评估

为了评估 GSO 中的个体 C 的性能，适应度函数包含了 MF 的最大化模糊项集支持度和语义可解释性。

$$fitness(C)=\sum_{x\in L_1}support(x)*\mathrm{GM3M} \tag{10}$$

其中：L_1 为 C 中基于 MF 获得模糊 1-项集的集合；$support(x)$ 为模糊 1-项集 x 的支持度；$GM3M$ 是量化 MF 可解释性的一种度量。$GM3M$ 被定义为三个度量的几何平均数，其值在 0(可解释性最低水平)和 1(可解释性最高水平)之间。定义为[14]

$$GM3M=\sqrt[3]{\delta\cdot\gamma\cdot\rho} \tag{11}$$

其中：δ、γ 和 ρ 为三个互补的度量标准，分别表示 MF 的位移(δ)、MF 的横向支持度(γ)和 MF 的面积相似度(ρ)。

如果模糊项集的支持度大于用户定义的最小支持度阈值($minSup$)，则认为是模糊项集。模糊项集 x 的支持度定义如下：

$$support(x)=\frac{count(x)}{|T|} \qquad count(x)=\sum_{t\in T}\mu_x(t) \tag{12}$$

其中：$|T|$ 为数据库 T 中的例子数量；$\mu_x(t)$ 为例子 t 与项集 x 的匹配程度。

3.3 举例分析

本文考虑一个精确的数据库 T 和一个不精确的数据库 $\tilde{T}$，$\tilde{T}$ 中包含了区间值混合和模糊的例子。两者都有两个变量(Age 和 $Height$)和三个例子。表 1 和 2 显示了这些数据库的例子。

表 1　精确数据库 T 的三个例子

ID	Age	$Height$
t_1	30	1.80
t_2	22	1.80
t_3	22	1.82

表 2　不精确数据库 $\tilde{T}$ 的三个例子

ID	Age	$Height$
$\tilde{t}_1$	21~31 之间	三角模糊集合 (1.77;1.79;1.89)
$\tilde{t}_2$	20~26 之间	1.80
$\tilde{t}_3$	22~23 之间	1.80~1.83 之间

将一个清晰的例子 t 与一个模糊的例子 $\tilde{t}$ 之间的相容程度定义为构成每个例子的变量成对值之间的相容性的最小值。如果数字不属于间隔区间，则数字与间隔的兼容性为 0，否则为 1。清晰数与模糊集之间的相容性是模糊集在清晰数上的隶属函数。例如，例子 $\tilde{t}_1$ 与 t_1、t_2 和 t_3 之间的兼容程度为

$$\begin{aligned} Compatibility(t_1,\tilde{t}_1)&=\min(1,0.9)=0.9 \\ Compatibility(t_2,\tilde{t}_1)&=\min(1,0.9)=0.9 \\ Compatibility(t_3,\tilde{t}_1)&=\min(1,0.7)=0.7 \end{aligned} \tag{13}$$

基于这个定义，精确数据库与不精确数据库之间的兼容程度由其成对示例之间的最小兼容性来定义。例如，数据库 T 和 $\tilde{T}$ 之间的兼容性为

$$\begin{aligned} Compatibility(T,\tilde{T})&=\min\begin{pmatrix} Compatibility(t_1,\tilde{t}_1), \\ Compatibility(t_2,\tilde{t}_2), \\ Compatibility(t_3,\tilde{t}_3) \end{pmatrix} \\ &=\min(0.9,1,1)=0.9 \end{aligned} \tag{14}$$

本文考虑模糊 1-项集 $x_1=(Age,Young)$，$Young$ 标签具

有的 MF：

$$\mu_{Young}(x)\begin{cases}1 & x<20\\ 2-x/20 & 20\le x\le 40\\ 0 & x>40\end{cases} \tag{15}$$

在精确的数据库 T 中，x_1 的支持度可以计算为 $support_T(x_1)=(0.5+0.9+0.9)/3=0.77$，但是这个值不能在不精确数据库中计算。

当一个不精确数据库由模糊值例子组成时，该项集的未知支持度就是一个模糊集合。形式为

$$\mu_{support_{\tilde T}(x)}(v)=\sup_{support_T(x)=v}\{Compatibility(T,\tilde T)\} \tag{16}$$

其中：在给定值 v 处的支持度的 MF，可表示为不精确数据 $\tilde T$ 与精确数据库 T 之间的最大兼容性，即 $support_T(x_1)=v$。如果问题被首先分成几个区间值问题，则支持度可以有效地计算。即设 $[\tilde t]_\alpha$ 是 $\tilde t$ 的一个 α_{cut}，表示一个区间。那么模糊支持的 α_{cut} 表示为以下区间：

$$[support_{\tilde T}(x)]_\alpha=\left[\begin{array}{l}\frac{1}{|\tilde T|}\sum_{\tilde t\in\tilde T}\min\{\mu_x(t)\,|\,t\in[\tilde t]_\alpha\},\\ \frac{1}{|\tilde T|}\sum_{\tilde t\in\tilde T}\max\{\mu_x(t)\,|\,t\in[\tilde t]_\alpha\}\end{array}\right] \tag{17}$$

例如，考虑模糊 1-项集 $x_2=(Height,Tall)$，标签 $Tall$ 的 MF 如式（18）所示。

$$\mu_{Tall}(x)\begin{cases}0 & x<1.50\\ (x-1.50)/0.40 & 20\le x\le 40\\ 1 & x>1.90\end{cases} \tag{18}$$

考虑一个简单的例子来说明适合度函数的计算。考虑一个工程问题，即直升机叶片上形成的冰与转速的相关性。表 3 显示了这个不精确训练数据库的例子。

表 3　这个数据库中的四个示例

ID	叶片上形成的冰	叶片转速
1	{[0.15,0.25]/下 +[0.75,0.85]/上}	三角模糊集(1, 3.3, 5.2)
2	8	0.18/上+0.85/下
3	[5.35, 6.5]	[4, 6.4]
4	(1, 1.1, 2.8)	[6.9, 8.2]

为简单起见，假设 MF 包含两个变量，结冰与速度是一致的，如图 3 所示。

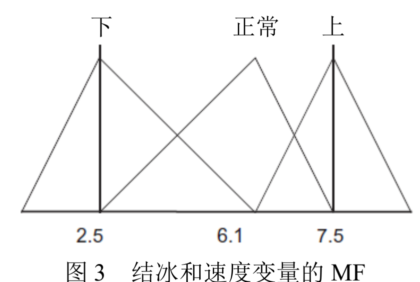


图 3　结冰和速度变量的 MF

设 $|\tilde A|$ 是模糊集 A 的平均值，它是一个区间。为了简化说明，表 4 中仅收集了不同分区中每个变量的支持度的平均值。假设最小支持度为 0.2。1-项集被粗体显示。

令

$$\delta_{down}=\frac{abs(5-6.1)}{5-2.5}=0.88 \tag{19}$$

因此 $GM3M$ 的值为

$$GM3M=(\delta\gamma\rho)^{1/3}=(0.88\times1\times1)^{1/3}=0.958 \tag{20}$$

那么

$$\begin{aligned}|support(x)|&=[0.47,0.612]+[0.205,0.56]+[0.22,0.315]\\&=[0.895,1.487]\end{aligned} \tag{21}$$

且

$$|Fitness|=[0.895,1.487]\times0.958=[0.857,1.426] \tag{22}$$

表 4　不同分区中每个变量的支持度

ID	结冰/下	结冰/正常	结冰/上	速度/下	速度/正常	速度/上
1	[0.15, 0.252]	0	[0.75, 0.85]	[0, 1]	[0, 0.75]	[0, 0.008]
2	0	0	1	[0.82, 0.85]	0	[0.15, 0.18]
3	0	[0.714,0.791]	[0.14, 0.6]	[0, 0.4]	[0.41, 1]	[0, 0.252]
4	[0, 1]	[0, 0.083]	0	0	[0, 0.428]	[0.76,1]
Σ	[0.15, 0.252]	[0.714, 0.874]	[1.895,2.45]	[0.82,2.25]	[0.41,2.178]	[0.91,1.46]
$\vert support_{\tilde T(x)}\vert$	[0.03,0.312]	[0.178,0.218]	**[0.472,0.615]**	**[0.203,0.543]**	[0.102,0.544]	**[0.221,0.318]**

4　基于 FFP-growth 算法的 FAR 挖掘

在这个阶段，将上述优化过程中获得的最优 MF 用来从不确定数据中挖掘有用的 FAR。为了避免运行时间过长，本文提出了一种新的数据挖掘算法，即基于 *FFP-growth* 算法的 *U-FFP-growth*，用于从不确定数据库中挖掘 FAR。该算法有效地利用频繁模式树结构来挖掘规则，以避免产生大量候选项集和数据库重复扫描。

考虑一个不确定的数据库 $\tilde{T}$，其中包含了 $|\tilde{T}|$ 个案例，$S=\{S_1,\ldots,S_n\}$ 是从优化过程中得到的最优 MF 集合。其中：n 为变量数目；$S_j=\{s_1,\ldots,s_m\}$ 表示第 j 个变量的语言标签集合；m 为语言标签的数量。本文将对 *FFP-growth* 算法进行扩展，使其能够从包含不确定数据的数据库中挖掘 FAR。过程如下：

a) 对于每个数据 $\tilde{t}_i, i=1,\ldots,|\tilde{T}|$，且对于每个变量 $j, j=1,\ldots,n$，将其值 $\tilde{v}_{ij}(\tilde{t}_i=\tilde{v}_{i1},\ldots,\tilde{v}_{in})$ 转换为模糊集。

b)对于每个语言学术语，根据式(17)计算其模糊支持度。对于每个变量 j，如果它的语言项的模糊支持度高于或等于 $minSup$，则它将被添加到 1-项集的集合 L_1 中。

c)创建一个头域表，对 L_1 按照模糊支持度进行降序排序。

d)根据排序再次扫描不精确的数据库，并根据 $s\tilde{upport}$ 的最大值，建立模糊的频繁模式树。

e)确定每个项集的模糊支持度，并产生一个所有项集的列表，用来验证 $support_{\tilde{T}}(x)\geq minSup$。

f)从每个项集中构建所有可能的关联规则。为了实现这一点，首先从每个 q-项集中生成可能的关联规则，其中 $q\geq 2$，项 $(x_1,x_2,\ldots,x_q)$ 表示为 $x_1\wedge\ldots\wedge x_{k-1}\wedge x_{k+1}\wedge\ldots\wedge x_q\rightarrow x_k, k=1,\ldots,q$。

然后本文计算规则的置信度，并根据用户确定的 $minConf$ 返回相关规则。

所提出的用来从不确定数据中优化合适 MF 和挖掘有用 FAR 的 *U-MFL-FAR* 算法总体过程如算法 1 描述。

算法 1　*U-MFL-FAR* 算法

输入：

一个包含 $|\tilde{T}|$ 例子的不确定数据库 $\tilde{T}$。

一组隶属函数 $S=\{S_1,\ldots,S_n\}$，$S_k=\{s_1,\ldots,s_m\}$ 表示第 k 个变量与模糊分区关联的语言标签集合。

预定义的 $minSup$。

预定义的 $minConf$。

用于近似 $\tilde{T}(\alpha_{cuts})$ 中项集支持度的切割数。

GSO 算法种群规模 N。

迭代次数。

阈值率 η。

输出：与一组最佳 MF 关联的模糊关联规则。

阶段 1：从不确定数据中通过优化过程获得最佳 MF。

a)生成具有 N 个个体的初始 GSO 种群。

b)评估个体适应度。对于每个个体：

将数据库中的每个不精确值转换为模糊集

计算每个模糊 1-项集 x_i 的支持度。

确定 x_i 是否属于 1-项集。

设置个体的适应度值。

c)初始化阈值 L。

d)生成下一代发现者、搜索者和游荡者：

e) 如果最好的发现者个体没有改变，那么 $L=L-(L_{initial}*\eta)$。

f)如果 $L<0$，重新构建种群并初始化阈值 L。

g)如果没有达到最大迭代次数，转到步骤 d)。

阶段 2：从不确定数据中挖掘模糊关联规则。

使用该组最好的 MF，通过扩展的 *FFP-growth* 算法，从给定的不确定数据库中挖掘 FAR。

5　实验及分析

5.1　不确定数据库

为了分析所考虑方法的性能，本文使用了一个真实世界的不确定性数据库：Adult[15]。选择一些数据构建本文实验中的数据库，其中案例数量为 5 200，数据中总变量数量为 150，数据中不准确的变量数量为 100。不确定数据的类型有精确值；区间值；一组可能的值；定义在标签集上的概率分布；标签上一个不精确的概率分布。

5.2　相关算法和参数设置

在这些实验中，本文将所提出的方法与现有的一种 FAR 挖掘算法进行比较，即文献[6]描述的用于不确定数据的模糊 Apriori 挖掘算法：*FApriori*。为了从不确定数据库中获得高质量的 FAR，该算法用不准确的值对输入数据进行建模，将每个输入值转换成一个模糊集合。

两种方法的参数如表 5 所示。对于本文方法，选择了在大多数情况下运行良好的标准通用参数，而不是特定值。其他方法的参数是根据相关文献中建议的值进行设定。

另外，所有实验中每个数据为 5 次运行的平均结果。Adult 数据库中的初始语言分区是由五个语言术语组成，以均匀分布的三角形 MF 形式构建，这些都是由系统专家预先定义。

表 5　各种方法的参数设置

算法	参数
FApriori	$\gamma=2$、$\alpha_{cuts}=7$、$minSup=0.3$ 和 $minConf=0.8$
U-MFL-FAR	$PopSize=50$、$N_{Eval}=200$、初始搜索角度 $\lambda^z=\pi/4$，最大搜索角度 $\omega_{\max}=\pi/a^2$，最大转向角 $\gamma_{\max}=\pi/2a^2$，常数 $a=\sqrt{s+1}$、$minSup=0.3$ 和 $minConf=0.8$

5.3　MF 学习过程中 α_{cuts} 的影响分析

在本节中进行了一些实验来分析置信区间 α_{cuts} 对 MF 学习的影响。为了使这个分析更容易解释，表 6 显示了在 3 个不同的 α_{cuts} 值(5、7 和 9)和 9 个不同的 $minSup$ 值(0.1 到 0.9)情况下的实验结果。其中：*fitness* 是适合度函数的值；*Support*

是 1-项集的平均支持度；$GM3M$ 是可解释性度量值；$\#L_1$ 是生成的 1-项集的数量。

通过分析获得的结果，可以看到 $fitness$、$Support$ 和 $\#L_1$ 随着切割次数的增加而增加，但在 9 次切割时增加量较低，且需要较长时间才能完成。因此，本文使用 7 次切割，在适应度值与 $GM3M$ 测度之间取得了很好的平衡。

表 6　具有不同 α_{cuts} 和 $minSup$ 值下本文 U-MFL-FAR 中 MF 优化的结果

minSup	$\alpha_{cuts}=5$				$\alpha_{cuts}=7$				$\alpha_{cuts}=9$			
	|Fitness|	|Support|	GM3M	#L1	|Fitness|	|Support|	GM3M	#L1	|Fitness|	|Support|	GM3M	#L1
0.1	[6.9,27.3]	[7.1,34.0]	0.82	43	[7.6,27.6]	[8.5,30.9]	0.93	48	[7.6,29.6]	[8.9,34.7]	0.94	50
0.2	[4.2,25.8]	[4.2,28.4]	0.91	36	[4.5,27.9]	[4.7,28.8]	1.0	36	[4.5, 28.1]	[4.5,29.3]	1.0	35
0.3	[4.4,11.2]	[4.4,17.1]	0.78	24	[4.8,12.4]	[7.8,20.1]	0.84	27	[5.1, 12.8]	[8.0, 24.5]	0.86	28
0.4	[1.3,12.5]	[1.3,13.1]	0.93	15	[1.5,13.9]	[1.6,14.1]	1.0	16	[1.5, 14.0]	[1.7, 14.0]	1.0	16
0.5	[0.3, 7.4]	[0.4, 8.4]	0.91	9	[0.6, 8.0]	[0.7, 9.2]	0.95	10	[0.6, 8.4]	[0.7, 9.6]	0.98	10
0.6	[3.2, 6.1]	[3.2, 6.9]	0.83	9	[3.8, 7.1]	[4.9, 9.2]	0.89	12	[3.7, 7.2]	[4.9, 9.5]	0.90	11
0.7	[2.1, 3.2]	[1.9, 4.8]	0.86	6	[2.3, 3.6]	[3.1, 4.9]	0.91	6	[2.4, 4.0]	[3.2, 5.1]	0.90	6
0.8	[2.2, 2.8]	[2.8, 4.8]	0.78	6	[2.4, 3.2]	[3.6, 4.9]	0.84	6	[2.2, 3.3]	[4.0, 4.8]	0.85	7
0.9	[0.8, 1.0]	[1.0, 1.4]	0.75	2	[1.0, 1.2]	[1.3, 1.7]	0.85	2	[1.1, 1.2]	[1.3, 1.7]	0.86	2

5.4　挖掘 FAR 的性能比较

首先，为了验证本文 MF 优化过程的有效性，将具备 MF 优化(U-MFL-FAR)和不具备 MF 优化(初始 MF)的挖掘 FAR 方法进行比较，结果如表 7 所示。可以看出，U-MFL-FAR 给出的适应度函数值比初始模糊分区给出的值要好，在平均支持度和可解释性度量 $GM3M$ 上也取得了很好的平衡。

这是因为 MF 的学习优化使本文可以更容易地从不精确数据中挖掘更相关的 FAR。另一方面，在不同 $minSup$ 值下，U-MFL-FAR 的 $GM3M$ 量度取得了很好的值，其对 MF 的原始形状作了优化调整，并保持了 MF 在合理水平上的语义解释能力。而基于初始模糊 MF 的挖掘中，总是获得 $GM3M$ 度量的最大值，这是因为这个度量致力于保持 MF 的原始形状，而且不被修改。

表 7　本文提出的 U-MFL-FAR 挖掘 FAR 的结果

minSup	U-MFL-FAR				初始 MF			
	$\|Fitness\|$	$\|Support\|$	$GM3M$	$\#L_1$	$\|Fitness\|$	$\|Support\|$	$GM3M$	$\#L_1$
0.1	[7.6, 27.6]	[8.5, 30.9]	0.9	48	[7.4, 25.5]	[7.4, 25.5]	1.0	39
0.2	[4.5, 27.9]	[4.7, 28.8]	1.0	36	[4.1, 26.0]	[4.1, 26.0]	1.0	32
0.3	[4.8, 12.4]	[7.8, 20.1]	0.6	27	[3.5, 8.2]	[3.5, 8.2]	1.0	15
0.4	[1.5, 13.9]	[1.6, 14.1]	1.0	16	[1.5, 13.3]	[1.5, 13.3]	1.0	11
0.5	[3.8, 7.1]	[4.9, 9.2]	0.8	12	[2.3, 4.6]	[2.3, 4.6]	1.0	6
0.6	[0.6, 8.0]	[0.7, 9.2]	0.9	10	[0.2, 5.7]	[0.2, 5.7]	1.0	6
0.7	[2.3, 3.6]	[3.1, 4.9]	0.7	6	[0.8, 1.7]	[0.8, 1.7]	1.0	4
0.8	[2.4, 3.2]	[3.6, 4.9]	0.7	6	[0.8, 1.2]	[0.8, 1.2]	1.0	3
0.9	[1.0, 1.3]	[1.3, 1.7]	0.7	2	[0.7, 1.2]	[0.7, 1.2]	1.0	2

然后，将本文方法与 $FApriori$ 方法在挖掘 FAR 性能方面进行比较，结果如表 8 和 9 所示。表 8 和 9 分别显示了当 $minConf$ =0.8 且 $minSup$ 为不同值时和 $minConf$ 为不同值且 $minSup$ =0.3 时，算法得到的规则的数量，其中也包含了使用初始模糊 MF 获得的规则。

表 8　不同 minSup 值下，各种算法获得规则的数量(minConf=0.8)。

minSup	U-MFL-FAR	$FApriori$	初始模糊分区
0.1	54	42	41
0.2	41	31	27
0.3	33	28	23
0.4	23	17	14
0.5	16	9	7
0.6	14	7	4
0.7	7	4	2
0.8	4	1	0
0.9	2	0	0

表 9　不同 minConf 值下，各种算法获得规则的数量(minSup=0.3)。

minConf	U-MFL-FAR	$FApriori$	初始模糊分区
0.1	423	415	415
0.2	328	312	313
0.3	257	246	245
0.4	172	167	167
0.5	107	104	103
0.6	69	65	65
0.7	44	38	38

0.8	33	28	23
0.9	5	4	4

分析这些结果可以看出，*U-MFL-FAR* 提取出了较多的 FAR，其数量大于或等于 *FApriori* 算法和初始模糊 MF 所获得的 FAR 的数量，从而获得一组合理的 FAR。其中，*FApriori* 算法使用模糊先验挖掘算法从初始模糊分区中挖掘 FAR，而本文使用了 *U-FFP-growth* 方法来挖掘初始模糊分区中的 FAR。显然，*U-FFP-growth* 算法比 *FApriori* 算法更有效，这是因为 *U-FFP-growth* 是基于频繁模式树结构。

因此可以得出结论，当不调整隶属函数时，规则的数量会减少。对于高支持度的规则，MF 优化操作的效果更加显着。例如，初始模糊分区中没有执行 MF 学习，则没有挖掘出支持度高于 0.8 的规则。

6 结束语

本文提出了一种新的不确定数据模糊规则挖掘算法，称为 *U-MFL-FAR*。通过自适应学习合适的 MF，最大限度地提高了 MF 的支持度和解释性度量，并提出了一种基于 FFP-growth 算法的新算法 *U-FFP-growth*，从不确定数据中挖掘有用的 FAR。通过实验结果表明，MF 的优化过程使本文能够挖掘出更加相关的 FAR。此外，使用可解释性度量 GM3M 可以避免原始 MF 发生较大的改变，在大多数情况下保留了原始的语义解释性。另外，采用了三元组语言表示来大大减少优化过程中的搜索空间。

参考文献：

[1] 高琳，赵书良，赵骏鹏，等．基于超图的关联规则可视化方法 [J]. 计算机应用研究, 2017, 34 (10): 2933-2937.

[2] 王爽，王国仁．面向不确定感知数据的频繁项查询算法 [J]. 计算机学报, 2013, 36 (3): 571-581.

[3] Kim J, Han M, Lee Y, *et al.* Futuristic data-driven scenario building: incorporating text mining and fuzzy association rule mining into fuzzy cognitive map [J]. Expert Systems with Applications, 2016, 57 (4): 311-323.

[4] Tazaree A, Eftekhari-Moghadam A M, Sajjadi-Ghaem-Maghami S. A semantic image classifier based on hierarchical fuzzy association rule mining [J]. Multimedia Tools & Applications, 2014, 69 (3): 921-949.

[5] 杨英杰，邱卫．基于时间衰减模型的模糊会话关联规则挖掘算法 [J]. 计算机应用研究, 2017, 34 (1): 128-131.

[6] Palacios A M, Gacto M J, Alcalá-Fdez J. Mining fuzzy association rules from low-quality data [J]. Soft Computing, 2012, 16 (5): 883-901.

[7] Li G Y, Li Y X, Li W, *et al.* Learning algorithm of parameters about fuzzy Membership functions based on the RBF neural network [C]// Proc of International Conference on Educational and Network Technology. Piscataway, NJ: IEEE Press, 2010: 282-286.

[8] 侯燕，刘辛．基于主从架构和 GA 的模糊关联规则挖掘算法 [J]. 控制工程, 2017, 24 (2): 276-282.

[9] Song A, Song J, Ding X, *et al.* Utilizing bat algorithm to optimize membership functions for fuzzy association rules mining [C]// Proc of International Conference on Database and Expert Systems Applications. Berlin: Springer. 2017: 496-504.

[10] Alcala R, Gacto M J, Herrera F. A fast and scalable multiobjective genetic fuzzy system for linguistic fuzzy modeling in high-dimensional regression problems [J]. IEEE Trans on Fuzzy Systems, 2011, 19 (4): 666-681.

[11] Chen C H, Hong T P, Lee Y C, *et al.* Finding active membership functions for genetic-fuzzy data mining [J]. International Journal of Information Technology & Decision Making, 2015, 14 (6): 1215-1242.

[12] 王飞，缑锦．基于多变异粒子群优化算法的模糊关联规则挖掘 [J]. 计算机科学, 2013, 40 (5): 217-223.

[13] Daryani N, Hagh M T, Teimourzadeh S. Adaptive group search optimization algorithm for multi-objective optimal power flow problem [J]. Applied Soft Computing, 2016, 38: 1012-1024.

[14] Safaee B, Mashhadi S K M. Fuzzy membership functions optimization of fuzzy controllers for a quad rotor using particle swarm optimization and genetic algorithm [C]// Proc of International Conference on Control, Instrumentation, and Automation. Piscataway, NJ: IEEE Press, 2016: 256-261.

[15] 陈爱东，刘国华，费凡，等．满足均匀分布的不确定数据关联规则挖掘算法 [J]. 计算机研究与发展, 2013, 50 (s1): 186-195.